Available online at www.sciencedirect.com
ScienceDirect
SoftwareX 5 (2016) 80–83
SoftwareX
www.elsevier.com/locate/softx

NCBI Mass Sequence Downloader – Large dataset downloading made easy

F. Pina-Martins[a,b,*], O.S. Paulo[a]

[a] *Computational Biology and Population Genomics Group, Centre for Ecology, Evolution and Environmental Changes, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal*
[b] *Departamento de Biologia e CESAM, Univ. de Aveiro, Portugal*

Received 15 October 2015; received in revised form 6 April 2016; accepted 28 April 2016

Abstract

Sequence databases, such as NCBI, are a very important resource in many areas of science. Downloading small amounts of sequences to local storage can easily be performed using any recent web browser, but downloading tens of thousands of sequences is not as simple.

NCBI Mass Sequence Downloader is an open source program aimed at simplifying obtaining large amounts of sequence data from NCBI databases to local storage. It is written in python (can be run under both python 2 and python 3), and uses PyQt5 for the GUI. The program can be run in either graphical or command line mode.

Source code is licensed under the GPLv3, and is supported on Linux, Windows and Mac OSX. Available at https://github.com/ElsevierSoftwareX/SOFTX-D-15-00072.git, https://github.com/StuntsPT/NCBI_Mass_Downloader

Keywords: Bioinformatics; Genomics; Molecular evolution; BLAST

Code metadata

Current Code version	3.1
Permanent link to code / repository used of this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-15-00072.git
Legal Code License	*GPLv3; Biopython license*
Code Versioning system used	*git*
Software Code Language used	*Python; PyQt5*
Compilation requirements, Operating environments & dependencies	*Python and (optionally for the GUI) PyQt5*
If available Link to developer documentation / manual	https://github.com/ElsevierSoftwareX/SOFTX-D-15-00072.git
Support email for questions	*f.pinamartins@gmail.com/frmartins@ciencias.ulisboa.pt*

Software Metadata

Current software version	3.1
Permanent link to executables of this version	https://github.com/ElsevierSoftwareX/SOFTX-D-15-00072.git
Legal Software License	*GPLv3; Biopython license*
Computing platform / Operating System	*GNU/Linux, Microsoft Windows, Mac OSX, any other where python and (optionally for the GUI) PyQt5 are available.*
Installation requirements & dependencies	*Python and (optionally for the GUI) PyQt5*
If available Link to user manual — if formally published include a reference to the publication in the reference list	https://github.com/ElsevierSoftwareX/SOFTX-D-15-00072.git
Support email for questions	f.pinamartins@gmail.com/frmartins@ciencias.ulisboa.pt

* Corresponding author at: Computational Biology and Population Genomics Group, Centre for Ecology, Evolution and Environmental Changes, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal.
E-mail address: f.pinamartins@gmail.com (F. Pina-Martins).

1. Introduction

National Center for Biotechnology Information (NCBI) sequence databases are nowadays a resource of unquestionable

http://dx.doi.org/10.1016/j.softx.2016.04.007

importance for researchers in many areas of science [1]. The current count of sequences available in this database as of 15 December 2015 ascends to over 18.9×10^7 sequences and 20.3×10^{10} base pairs (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt), representing roughly 2.5 Tb of compressed data and keeps growing. Due to advances in sequencing technology, the amount of sequence data required by investigators has increased by orders of magnitude in the last few years. This has naturally led to increased use of the NCBI databases by investigators for retrieving sequence data.

NCBI Mass Sequence Downloader provides a user friendly interface and automated error checking for downloading large sets of sequence data.

2. Problems and background

Although downloading sequences from NCBI can be done in a simple fashion using any standards compliant web browser via the *Entrez* [2] web portal, this method does not scale well, and downloading large amounts of sequences (in the order of the tenths of thousand) from these databases can cause problems when performed this way (https://www.biostars.org/p/43970/). Furthermore, manually performing this type of tasks is time consuming and error prone, which may hamper the reproducibility of scientific work.

For retrieving large sets of data, NCBI provides the *E-utilities* API [2], although it can be difficult to use by investigators without an IT related background, despite it's through documentation. Frameworks exist, written in various popular languages, such as *Python*, *Perl* or *Ruby*, that provide some level of abstraction for using this API, such as *Biopython* [3], *BioPerl* [4], or *BioRuby* [5], respectively. However, these too, require some degree of programming knowledge to use, rather than providing end-user packages ready to use for a specific purpose. This leaves investigators without a simple, ready-made solution. Although this is not much of a problem for someone with a bioinformatics background, it poses a serious issue for someone with a molecular biology background, who may frequently require this kind of data, but lack the programming skills to use one of the mentioned frameworks or the API.

By using NCBI's API, our program intends to solve the problem of retrieving large datasets, in a user friendly, automated, and reproducible way. The tool is therefore aimed at molecular biologists that do not have an IT related background, but need to download large datasets from the NCBI databases.

3. Software framework

3.1. Software architecture

NCBI Mass Sequence Downloader is written in python (http://www.python.org) and can be run under both python 2 and python 3. The command line interface (CLI) version of the program can be run on any OS that has python available. The GUI version further requires PyQt5 (http://www.riverbankcomputing.com/software/pyqt/intro) available, which means all major currently used operating systems such as GNU/Linux, MS Windows and Mac OSX are supported. The program uses a slightly altered (changed the *import* statements) module from *Biopython* [3] to *Entrez* [2], which is included with the software. This avoids the need to have *Biopython* installed, which, despite being a popular library in the bioinformatics community, is not usually so for molecular biologists. The consequence of this convenience for the user is a higher maintenance requirement since it makes it necessary to keep up with the upstream *Entrez* module. However, this *Biopython* module has not had many recent changes, and merging them into *NCBI Mass Sequence Downloader* has so far been trivial.

The program consists of essentially three modules — a back-end, a front-end and the *Entrez* module. If the program is run without arguments, the GUI version is launched (Fig. 1), but if the program is run with arguments, the command line version will be run instead. This makes the program quite flexible to use in different environments.

The program's source code is available on github (https://github.com/ElsevierSoftwareX/SOFTX-D-15-00072.git), along with binaries for GNU/Linux, MS Windows and Apple OSX.

3.2. Software functionalities and limitations

NCBI Mass Sequence Downloader is made to solve a single task — downloading sets of sequences from the NCBI databases. For this, the user should provide an email address for eventual contact from NCBI (which is sent only to NCBI), the database to be queried, the search query, and a path to the file for the downloaded sequences. Download progress is indicated in both user interfaces.

Currently, the program is limited to downloading sequences in the FASTA format and to NCBI databases, but data from several databases can be retrieved: *nucleotide, nuccore, nucgss, protein, genome* and *popset.*

3.3. Internal routines and error handling

Once the program is requested to start the download, it queries the selected NCBI database for the inputted search term. It will then store the returned sequence IDs in memory, and begin downloading the respective records in batches of 3000 sequences. Every batch is temporarily stored in memory, and once 3000 sequences are downloaded, they are immediately stored in the output file, flushed from memory, and then, the next batch is processed.

After all the records are downloaded, the output file is parsed and its sequences' IDs matched to the originally retrieved sequence IDs. If any sequences are missing, a new pass is made to retrieve them. This process is repeated as often as necessary until all the requested sequences are stored in the output file. Stopping the program at any time will not affect any sequences already stored in the output file.

If a pre-existing FASTA file is selected as the output file, instead of overwriting it, the file is parsed, and the sequences' IDs are retrieved and compared to those returned by NCBI for the requested query. Any already present sequences are not downloaded again, and any missing sequences are appended to the

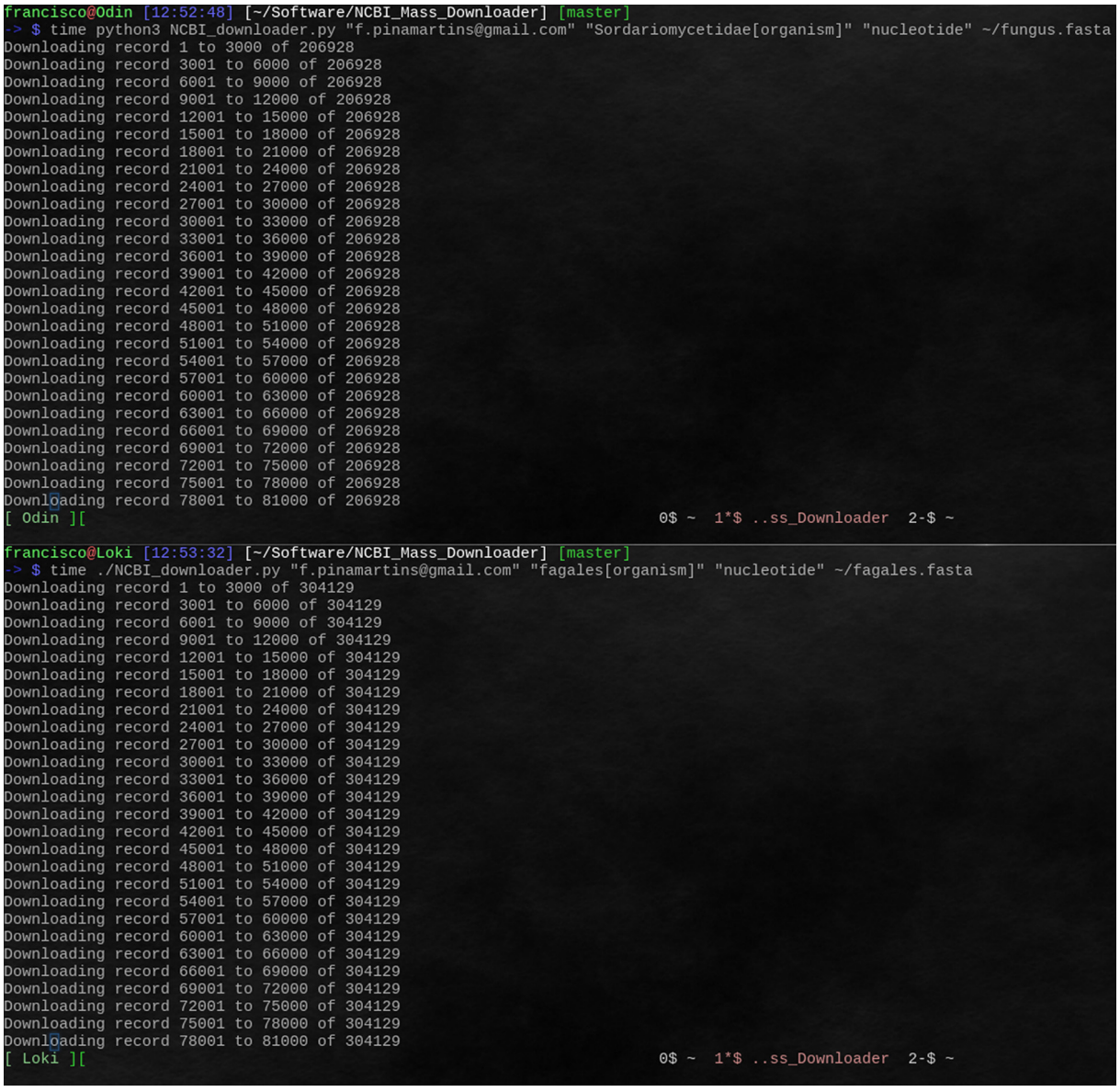

Fig. 1. A screenshot of *NCBI Mass Sequence Downloader* running under a GNU/Linux based OS in command line interface, downloading sequences matching the queries "Sordariomycetidae[organism]" (above) and "Fagales[organism]" (below).

end of the file. This capability makes it possible to resume any canceled download.

NCBI Mass Sequence Downloader will handle any server errors thrown during sequence retrieval by pausing all activity for eight seconds and then retrying. Five such failures in a row, cause a further 20 s pause before trying a retrieval operation again.

3.4. Future plans

Several developments are expected for future releases of *NCBI Mass Sequence Downloader*, such as being able to get data in formats other than FASTA, adding an online interactive help system to the GUI or even the capability to query databases other than NCBI. We expect to keep the software maintained to work with future versions of python, Qt, and database APIs for the foreseeable future.

4. Illustrative examples

4.1. Example use case

A molecular biologist has to analyze a hypothetical dataset of transcriptomic data of a plant–fungus system (*Castanea dentata*, *Cryphonectria parasitica*). In order to identify which sequences can be considered "plant" and which can be considered "fungus", instead of downloading the entire "nt" database from NCBI and running BLAST [6] queries against it, by using *NCBI Mass Sequence Downloader*, it is possible to download only the sequences of the *Fagales* (plants) order and *Sordariomycetidae* (fungus) subclass, and run the required BLAST queries against the resulting files. This would considerably reduce both download and query time, provide the user with more specific results and enable a simpler downstream data filtering process.

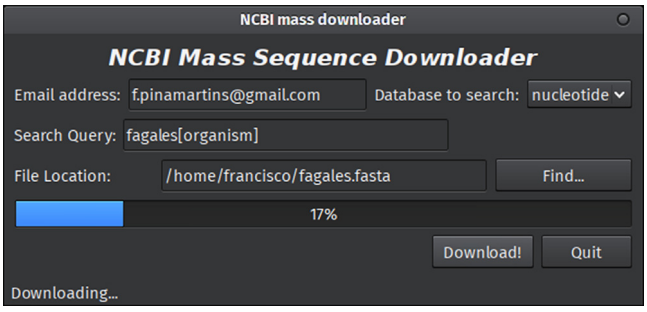

Fig. 2. A screenshot of *NCBI Mass Sequence Downloader* running under a GNU/Linux based OS in graphical user interface, downloading sequences matching the query "Fagales[organism]".

An example study where *NCBI Mass Sequence Downloader* could have been useful is [7], where the investigators performed BLAST searches against several datasets that could have been quickly obtained and segregated with this software.

4.2. The command line interface (CLI)

In order to use the CLI version of the program for solving the problem described in Section 4.1, the user needs to run the program with the following arguments: "user email address", "query term", "database to query" and "output file". Screenshots of *NCBI Mass Sequence Downloader* performing this specific task in the CLI environment can be seen in Fig. 1.

In this test, the plant query took ∼48 min to download all 304,129 records, roughly 1.2 GB of sequence data. The fungi query took ∼42 min to download all 206,928 records, amounting to approximately 2.0 GB of sequences.

4.3. The Graphical User Interface (GUI)

In order to use the GUI version of the program, the user needs to run the program without any arguments. This will bring up the interface main window (Fig. 2), where the user can enter the required information to proceed with the downloading of the queried sequences.

The performance of the GUI version was essentially the same as the one obtained using the CLI method.

4.4. Using the alternative methods

Using the *Entrez* web portal to download the sequences mentioned in the example resulted in having to attempt each of the downloads several times until all the sequences were obtained. Furthermore, the download would simply stop without issuing any error messages, and it was thus, necessary to manually verify that all the requested sequences had been downloaded (which did not happen in the first three tries for the plant dataset and for the first two times for the fungus dataset). This method was thus more time consuming and required manual user intervention several times until all the requested data was locally stored.

The E-Utilities API can also be used directly. In order to get the example data using this method, the following actions need to be taken:

1. Make the search query to the NCBI servers.
2. Retrieve the "Query Key" and "WebEnv" variables.
3. Request the sequences in blocks of up to 10^4 until all are downloaded.

This can be done manually, but it is a tedious and error prone process (step 3 would have to be performed 62 times to download all sequences from the example case). Alternatively, this behavior can be scripted to automate the process, but that requires programming skills, which may act as a barrier to molecular biologists.

5. Conclusions

Although querying the NCBI database and downloading the respective sequences can usually be done from the web browser, when it is necessary to download large amounts of sequences, this procedure becomes unreliable since the probability of download problems increases with its size and the *Entrez* web portal does not provide a way to resume interrupted downloads. Using the alternate method — via the *E-utilities* API requires programming skills and not every molecular biologist is equipped to deal with that. These issues make the process of retrieving large datasets from NCBI an error prone and attention demanding process, unless the user has some programming skills.

NCBI Mass Sequence Downloader was designed to fill in this gap. To allow anyone without programming skills to easily download large sequence datasets from the NCBI databases, in an automated, reliable and reproducible way.

Furthermore, the possibility to choose the interface, makes *NCBI Mass Sequence Downloader* appropriate to use both on desktop and on the command line based systems.

Acknowledgments

This study was financed by Portuguese National Funds, through FCT — Fundação para a Ciência e a Tecnologia, within the projects UID/BIA/00329/2013, SOBREIRO/0036/2009 and SFRH/BD/51411/2011.

References

[1] Miller H, Norton CN, Sarkar IN. BMC Res Notes 2009;2:101.
[2] Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J. Nucleic Acids Res 2010;38:D5.
[3] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Bioinform Oxf Engl 2009;25:1422.
[4] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. Genome Res 2002;12:1611.
[5] Goto N, Prins P, Nakao M, Bonnal R, Aerts J, Katayama T. Bioinforma Oxf Engl 2010;26:2617.
[6] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. J Mol Biol 1990;215:403.
[7] Haçariz O, Akgün M, Kavak P, Yüksel B, Sağiroğlu MŞ. BMC Genomics 2015;16:366.